D400–D412 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 2 December 2011*
doi:10.1093/nar/gkr1132

PubChem's BioAssay Database

Yanli Wang*, Jewen Xiao, Tugba O. Suzek, Jian Zhang, Jiyao Wang, Zhigang Zhou, Lianyi Han, Karen Karapetyan, Svetlana Dracheva, Benjamin A. Shoemaker, Evan Bolton, Asta Gindulyte and Stephen H. Bryant*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA

Received September 9, 2011; Revised November 7, 2011; Accepted November 8, 2011

ABSTRACT

PubChem (http://pubchem.ncbi.nlm.nih.gov) is a public repository for biological activity data of small molecules and RNAi reagents. The mission of PubChem is to deliver free and easy access to all deposited data, and to provide intuitive data analysis tools. The PubChem BioAssay database currently contains 500 000 descriptions of assay protocols, covering 5000 protein targets, 30 000 gene targets and providing over 130 million bioactivity outcomes. PubChem's bioassay data are integrated into the NCBI Entrez information retrieval system, thus making PubChem data searchable and accessible by Entrez queries. Also, as a repository, PubChem constantly optimizes and develops its deposition system answering many demands of both high- and low-volume depositors. The PubChem information platform allows users to search, review and download bioassay description and data. The PubChem platform also enables researchers to collect, compare and analyze biological test results through web-based and programmatic tools. In this work, we provide an update for the PubChem BioAssay resource, including information content growth, data model extension and new developments of data submission, retrieval, analysis and download tools.

INTRODUCTION

PubChem (1–3) (http://pubchem.ncbi.nlm.nih.gov) is a public information resource for archiving chemical structures and biological properties of small molecules and siRNA reagents. It is hosted by the National Center for Biotechnology Information (NCBI) (4), a division of the National Library of Medicine under the National Institutes of Health since 2004. The PubChem BioAssay database currently consists of bioactivity information generated by high-throughput screenings and medicinal chemistry studies. In addition, the PubChem BioAssay database contains a few dozen high-throughput RNAi screens against complete genomes. These data are integrated with the rest of the NCBI resources, making PubChem a widely used public information system for chemical biology and drug discovery research.

The PubChem BioAssay database is organized as a set of relational databases deployed on Microsoft SQL servers. The infrastructure allows for seamlessly storing the submitted bioassay records, tracking and versioning subsequent updates, and supporting data retrieval and analysis. PubChem provides a user-friendly deposition system to facilitate data exchanges and submissions. To make the vast bioactivity information easily accessible to the scientific community, PubChem provides a suite of integrated services enabling users to analyze biological test results, identify and validate drug targets and evaluate chemical and RNAi probes. To meet the increasing demand from public users and from rapid growth of data volume and complexity, PubChem maintains and develops its service to the community as a public data repository by optimizing and expanding its bioassay data model for supporting broader types of information, by developing infrastructure to ensure database scalability, by improving deposition system to ease information exchange, and by enhancing search, retrieval, analysis and download tools. In this work, we will provide brief descriptions of these important components of the PubChem BioAssay resource with an emphasis on the new developments within each section.

BioAssay DATA CONTENT

The PubChem BioAssay database currently contains 500 000 bioassay records, a total of over 130 000 000

*To whom correspondence should be addressed. Tel: +1 301 435 7811; Fax: +1 301 435 7793; Email: ywang@ncbi.nlm.nih.gov
Correspondence may also be addressed to Stephen H. Bryant. Tel: +1 301-435-7792; Fax: +1 301-435-7793; Email: bryant@ncbi.nlm.nih.gov

bioactivity summary results and 740 000 000 data points. These test results represent rich biological properties for 120 chemical probes, 1 600 000 small molecules and 60 000 RNAi reagents. They also provide unique annotations for over 5000 protein targets tested in small molecule screenings and 30 000 gene targets tested by RNAi screenings. This information is contributed by over 40 organizations including US government agencies, NIH-funded screening centers, pharmaceutical companies and worldwide research laboratories. A summary of bioassay depositions is available at http://pubchem.ncbi.nlm.nih.gov/sources#assay.

PubChem continues to host screening data generated by the NIH Molecular Libraries and Imaging Program (MLP) (http://commonfund.nih.gov/molecularlibraries/). Contributions from some other organizations were described previously (1). In the past 2 years, PubChem has additionally received bioassay depositions from organizations with large-scale screening facilities and individual research groups. For example, researchers at UCLA deposited their research data from a study on the survival of human embryonic stem cells with a chemical genomics approach (5). The ICCB-Longwood/NSRB Screening Facility at the Harvard Medical School (http://iccb.med.harvard.edu/) contributed a number of high-throughput screening data sets containing inhibition activity of small molecules for several biologically important targets. The Drosophila RNAi Screening Center (DRSC) (6) has recently submitted a number of data sets with RNA interference screening results against the *Drosophila* genome. As an endeavor to fight malarial disease (7) and share the scientific discovery with the worldwide community, Glaxo Smith Kline teamed up with public domain data providers and deposited the anti-malarial data sets into PubChem (http://www.prnewswire.com/news-releases/gsk-and-online-communities-create-unique-alliance-to-stimulate-open-source-drug-discovery-for-malaria-94430694.html). This exemplary model was followed by scientists at Abbot Labs: two data sets of bioassay data reporting collective information on a drug–target network study were submitted to PubChem (8). In addition, a cell viability assay and a caspase 3/7 assay were deposited by the Laboratory of Environmental Genomics at the Carolina Center for Computational toxicology, University of North Carolina at Chapel Hill (http://comptox.unc.edu/); and a USP14 inhibitor assay was deposited by the Finley and King Labs at the Harvard Medical School (9). Following the deposition of the siRNA circadian assay (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1904) contributed by collaborators, the Kay laboratory at the University of California at San Diego contributed two small molecule screening data sets to PubChem (10,11). These three data sets on circadian research are linked to each other in PubChem as related bioassays specified by depositors. This linkage makes an excellent showcase to demonstrate that RNAi screenings complement small molecule assays as they can be joined together to explore the key genes and proteins critical to the circadian pathway. Most of these depositions have accompanying results published in scientific journals, thereby giving PubChem a valuable role as a hub linking raw and annotated scientific data with their respective research papers.

PubChem collaborates with the European Bioinformatics Institute (EBI) and mirrors the full ChEMBL database (12). This data set covers over 30 000 publications from 17 scientific journals. PubChem assigns a unique PubChem BioAssay accession (AID) to each of the imported bioassay records, and provides cross-links to the respective ChEMBL web pages. Protein type targets with high annotation confidence from ChEMBL are recorded in PubChem as bioassay targets after mapping to GenBank records, while those with lower confidence are recorded as descriptive comments. PubChem flags the endpoints in ChEMBL, such as IC50, EC50, Ki, with the 'active concentration' attribute, and converts readout values to uM (micromolar) units in compliance with the PubChem data standard. This allows PubChem to link each ChEMBL assay to a subset of compounds with potency of $\leq$1 uM and $\leq$1 nM, respectively. For ChEMBL bioassay records that are derived from the same research article, PubChem tracks their inherited relationship and marks these assays as same-publication-based related bioassays, which could be useful from a research point of view. For additional context, PubChem retrieves from PubMed the title and abstract of the publication associated with a ChEMBL bioassay, presents them in the PubChem bioassay summary page and indexes them in Entrez to facilitate database search. PubChem further integrates ChEMBL data to the rest data content in PubChem with a set of data analysis tools enabling researchers to compare recently generated HTS data to research results reported in the literature to accelerate discovery process.

PubChem BioAssay STANDARD AND DATA MODEL

PubChem is aimed to accommodate diverse bioactivity information with a flexible BioAssay data model and database schema, and continues to expand the types of data it accepts as experimental methodologies evolve. The data model (1) was designed to unambiguously represent bioassay protocols, molecule target information, other cross-database references, bioactivity summary results and user-defined readout types associated with tested reagents. A depositor is able to provide as many detailed test results as needed for delivering the research findings. As a result, one can report replications of a specific readout as well as one or multiple series of dose–response data points. The PubChem bioactivity summary result, e.g. bioactivity outcome, score and active concentration attribute, allows one to rank and evaluate the hits identified in the screening experiment. It supports cross-links from the bioassay record to chemical probes, bioactive compounds, as well as a subset of compounds with certain potency. Moreover, it allows PubChem to provide tools to enable in-depth data analysis and comparison across multiple bioassay results.

PubChem provides several schemes for depositors to report targets for tested reagents. The 'classical' model for

a PubChem bioassay is a single protein target tested with the entire set of small molecule compounds. The 'panel' model reports multiple bioactivity outcomes against different targets as well as multiple cell lines or species. It can also be used to report results from multiple but highly related experiments. Through this model, depositor may designate the respective targets to a group of test results (TID), for example. Another model is to accommodate screening results of different types of test reagents, such as RNAi probe molecules. In this case, each tested RNAi reagent aims for its own target. Thus, a depositor can specify the gene target for each reagent and provide this information under a gene target column in the test result data table. Similarly, PubChem now allows other types of cross-reference, such as those to PubMed, GenBank or NCBI Probe databases, to be specified per each tested substance.

In addition to the bioassay relationship annotation provided by depositors, PubChem provides several computational methods to associate bioassays, resulting in several kinds of related bioassays based on target sequence similarity, common active compounds, common biological pathways as well as data abstracted from the same publication. As a result, RNAi screenings in PubChem are automatically linked to small molecule assays if the biologically responsive genes from an RNAi screening and the protein targets of small molecules are involved in the same pathway.

PubChem allows depositors to provide updates to their records. The nature of bioassay update could vary from fixing a simple typo to providing additional descriptive information, cross references, or test results. As an important aspect of a public archival system, the PubChem BioAssay data model and database schema are carefully designed with infrastructure for supporting, tracking and storing updates and modifications to the existing bioassay records. By default, PubChem shows and distributes the information from the most recent version of a bioassay record through its web services and FTP sites; however, earlier versions of the record can be retrieved through the BioAssay Summary web service upon user request.

Depositors may provide cross-references in their submissions to link the bioassay record to taxonomy, gene or 3D structure of the target. A new field recently added to the PubChem BioAssay data model allows one to highlight primary citations out of a list of PubMed cross-references. PubMed citations designated as primary should reference papers containing experimental information directly relevant to the bioassay record, thus helping PubChem users better understand the assay results. Reciprocally, such PubChem/PubMed direct links will also allow PubMed users to immediately access assay results, hence facilitating information integration by PubChem and other NCBI resources.

PubChem provides a generic bioassay data model to capture common elements essential for recording screening results. With the increasing growth in data diversity and request for recording information relevant to a specific project, a new data field, e.g. 'categorized comment', has recently been introduced, which allows depositors to provide information categories and textual data associated with each category. This mechanism offers the means and flexibility for depositors to provide the information pertinent to a focused research area, to comply with recommendations on data standard from a working group or to meet the guidelines of data exchange and sharing as required by a research community or consortium. This mechanism makes it easier for depositors to validate information prior to submission, and enables them to adequately describe projects for internal requirements. Meanwhile, such a semi-structured data model allows PubChem to accommodate a greater diversity of information content critical to multiple research communities. It also allows PubChem to tailor its tools to search, present and classify the information in the future. An example of a bioassay containing categorized comments is available at http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=540333.

All information from a BioAssay deposition can be wrapped up as an XML or ASN.1 data object. A complete list of data fields for the PubChem BioAssay data model, and detailed descriptions of their usage, can be obtained by following the XML schema or equivalent ASN.1 specification at the PubChem FTP web sites:

ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem.xsd
ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem.asn

PUBLIC ACCESS AND SEARCH

An individual record in the PubChem BioAssay database can be accessed directly through the BioAssay Summary service at http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=myAID, where 'myAID' is a valid numeric PubChem BioAssay accession (AID). This service provides access to all versions of deposited assay information, such as assay protocol, test result descriptions and data (Figure 1). It allows one to retrieve, view, and download test results through the 'Show Data' links. The service also lists information about the assay target, including depositor-provided molecular information and annotations derived by PubChem about protein family classification, the corresponding gene, pathway and homologous 3D structures. Furthermore, the BioAssay Summary service provides a central entry point to a set of data analysis tools for the bioactive compounds identified in the assay. These analysis tools can be accessed through the 'BioActivity Summary', 'Structure–Activity Analysis' and 'Structure Clustering' links, and allow one to cluster the scaffolds of the tested compounds, examine and visualize SAR relationships, and evaluate target specificity or promiscuity properties of the compounds. In addition, the 'Related BioAssays' section lists assays that may be related to the one under review and links to further detailed summary over the bioassay relationship. Cross-references to other NCBI databases, such as PubMed, are listed under the 'Links' section.

PubChem BioAssay records can be searched in the NCBI information retrieval system Entrez. Descriptive information content in the BioAssay database is indexed

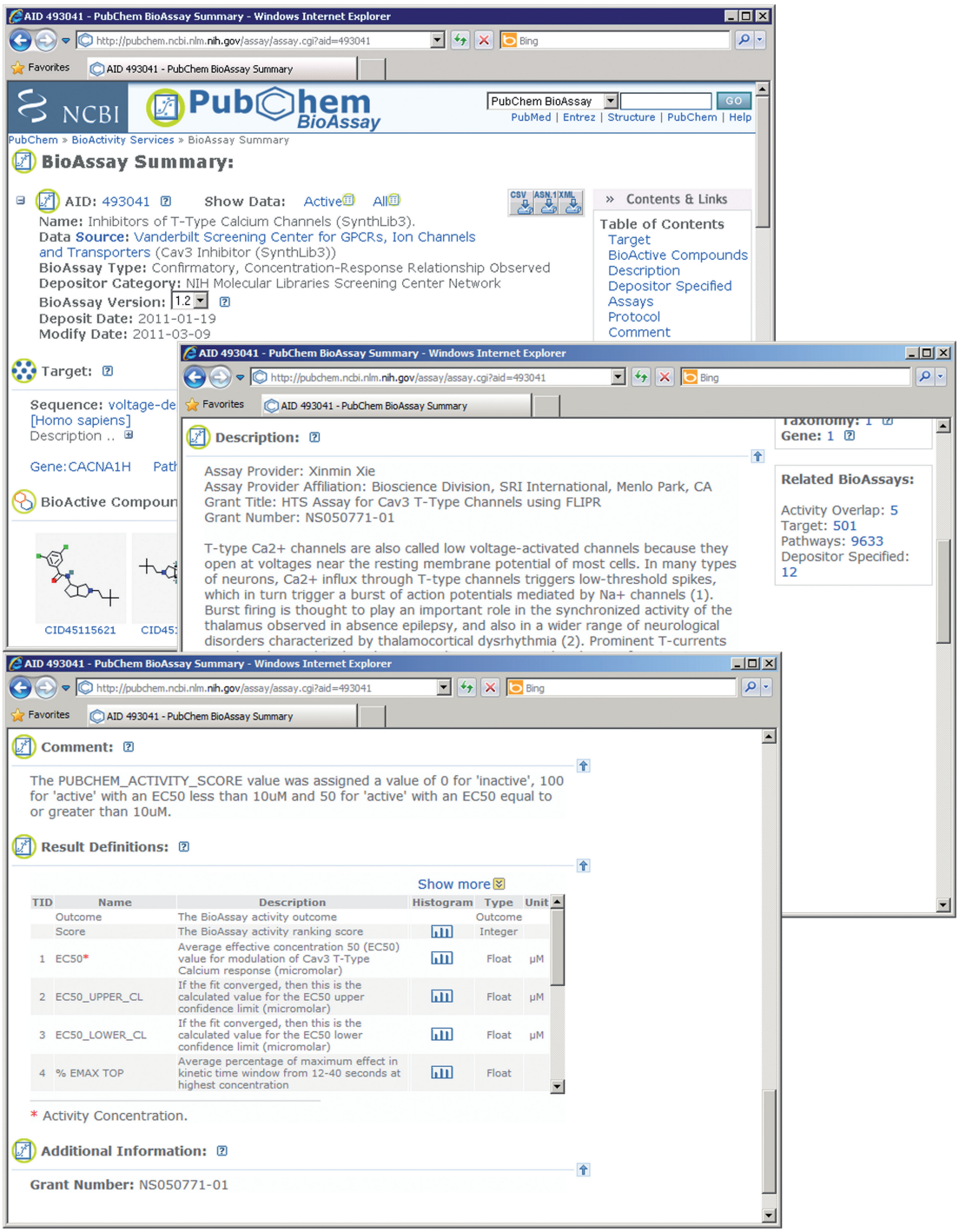

Figure 1. The summary view of a PubChem bioassay record. Assay results can be retrieved through the Show Data | Active and Show Data | All links.

under multiple fields to facilitate general as well as specific searches for bioassay records. A full list of indexed fields and filters, such as assay name, description, protocol, target description, readout name and tested chemical name, are documented at the PubChem Help page (http://pubchem.ncbi.nlm.nih.gov/help.html#PubChemindex). Entrez BioAssay search can be accessed at http://www.ncbi.nlm.nih.gov/pcassay/ or at http://pubchem.ncbi.nlm.nih.gov. Documentation for general use of the NCBI Entrez system is available at http://www.ncbi.nlm.nih.gov/Database/index.html. Detailed description for making an effective Entrez bioassay query has been described previously (1,3) and will be further described below.

The Entrez system allows users to perform a search with a single keyword, or a complex query against one or

multiple indexed fields. For example, one may use http://www.ncbi.nlm.nih.gov/pcassay?term=gleevec to find out assays containing the text word 'gleevec' in any part of the bioassay records. One can further limit the search to a specific data field, for example, to use http://www.ncbi.nlm.nih.gov/pcassay?term=gleevec [SynonymTested] to find out assays where the compound gleevec is tested, and use http://www.ncbi.nlm.nih.gov/pcassay?term=gleevec[SynonymTested] AND gefitinib[SynonymTested] to retrieve assays which tested both gleevec and gefitinib. For specific searches, one may use the Entrez 'Limits' page at http://www.ncbi.nlm.nih.gov/pcassay/limits. Users can pick a search field given in the 'Search Field Tags' section or build up a complex query using other input boxes on the page. One can further join the results from multiple searches using the 'Search History' features accessible from the 'Advanced' page at http://www.ncbi.nlm.nih.gov/pcassay/advanced.

In addition to indexing data submitted by depositor, PubChem also indexes data derived by PubChem, thus making bioassay data records more discoverable. One of the derived indexes is the gene name corresponded to the protein target, which includes official symbol and gene synonyms provided in the NCBI Gene database. Search through gene symbol name of the bioassay target can be advantageous as it may bring up assays which contain variations of protein target names and molecular identifiers. For example, one can collect over 600 assays targeting on human ERG following the query http://www.ncbi.nlm.nih.gov/pcassay?term=KCNH2[genesymbol], among which users can find a bioassay deposited by PDSP, a dozen hERG inhibitor screening assays from the Johns Hopkins Ion Channel Center, and several hundred bioassay records from ChEMBL. These bioassay depositions provide protein target references to a list of different protein records in Entrez Protein database. However, despite differences in protein identifiers, these records contain identical protein sequence information and point to the same gene record in the NCBI Gene database; hence, a larger set of bioassay records can be identified with a single query by gene name rather than by protein name or molecular identifier.

As the majority of bioassay records in PubChem are associated with publications, PubChem now indexes journal name and publication date for all cross-references to PubMed. Thus, one can search bioassays by limiting the query to the 'JournalName' field. One can also use the 'Cited Publication' menu on the Limits page to search assays associated with a selected journal. PubChem extended Entrez's 'auto-complete' feature to the BioAssay database which covers several index fields including 'JournalName', 'ProteinTargetName', 'TaxonomyName' and data 'SourceName'. This feature can be accessed on the 'Advanced' page, where selecting the 'JournalName' field and entering 'med' in the 'Search Builder' input box will bring up a list of journal names including 'Journal of Medicinal Chemistry', 'European Journal of Medicinal Chemistry' etc., for example.

Entrez presentation (e.g. DocSum report) for the PubChem BioAssay database has recently been converted to a display style generic to all Entrez databases (Figure 2). The DocSum report links to the full summary of a bioassay record through the bioassay title, and connects to the bioassay data table through the 'All data' or 'Active data' link. One can also access information related to a bioassay record following the 'Protein Target', 'Compounds, Active', 'PubMed Citation' links. New features have been developed for the BioAssay DocSum report to allow one to easily refine the search results and subsequently focus on a subset of assays of interest. BioAssay records among the search results are grouped and summarized under the 'Refine your results' section based on bioassay target, bioactivity potency, experiment type and depositor category. Using that interface, users can narrow down to a subset of assays targeting on a specific protein, or assays containing inhibitors of $\leq$1 uM (or $\leq$1 nM, not shown in Figure 2) potency. Furthermore, users can now download selected bioassay records using the 'BioAssay Download' function given in the 'Actions on your results' section. The 'Find related data' feature links to the cross-references among the resulted records to other NCBI databases.

BioActivity ANALYSIS TOOLS

PubChem provides web-based and programmatic tools for users to retrieve, analyze and export bioactivity data in the BioAssay database. A list of web-based bioactivity analysis tools and their URLs are summarized in Table 1, which can also be accessed from the PubChem web page at http://pubchem.ncbi.nlm.nih.gov/assay. These tools allow users to retrieve assay descriptions and data, review related bioassays, compare bioactivity data from multiple experiments and explore structure–activity relationship. Some of these tools have been described in detail previously (2). Enhancements for the data analysis tools and new services developed in the past 2 years are summarized below.

Compound-centric and target-centric BioActivity Summary tools

The BioActivity Summary service is a central tool for summarizing available bioactivity information for one or a set of compounds, bioassays and targets. The assay-centric view (Figure 3A) of this service presents the bioactivity data from the assay point of view as previously described (2). Two new components, e.g. compound-centric view and target-centric view, have been added to this service. The compound-centric page provides an overview of the bioactivity data available from the compound point of view. It summarizes the bioassay experiments and protein targets for each compound and groups such information based on bioactivity outcome and potency range (Figure 3B). Each summary count shown in the table also represents a link which, if followed, leads to detailed bioactivity results for the compound and the associated assays/targets. Most of the high-throughput screen data sets in PubChem contain bioactivity outcome specification, e.g. active versus inactive.

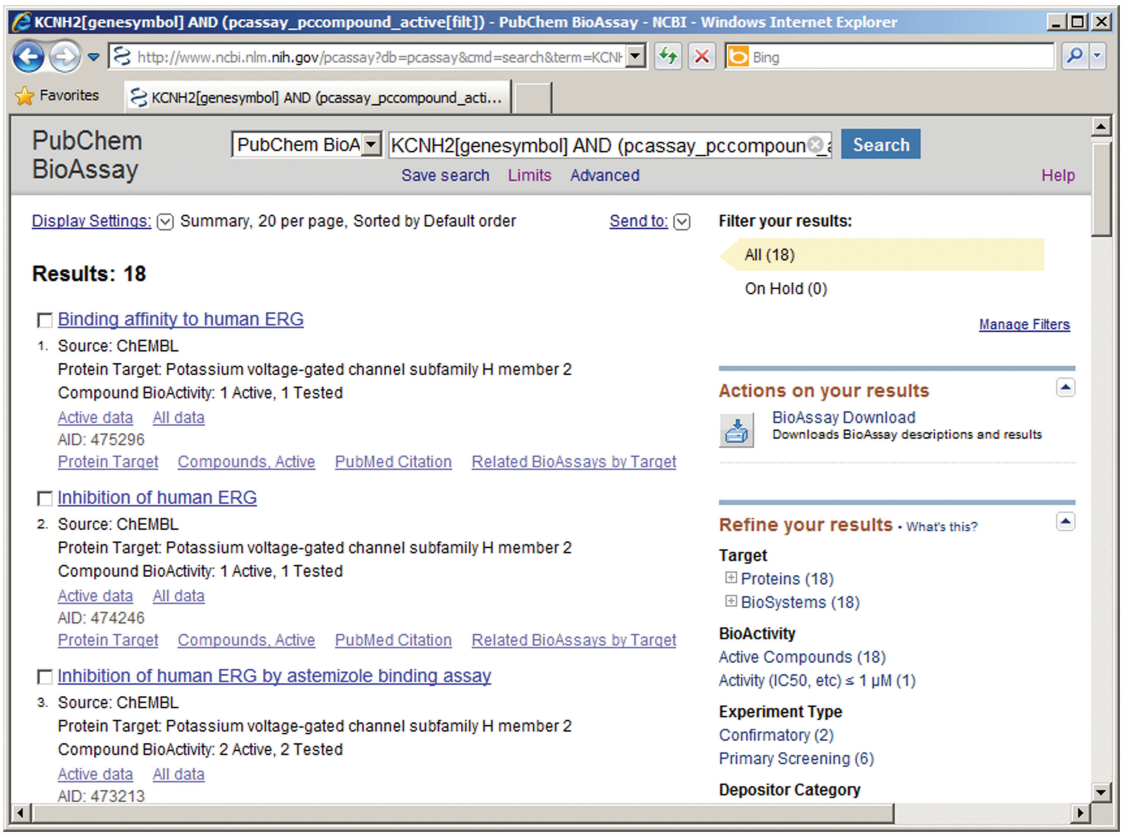

Figure 2. Entrez DocSum report for bioassay search results.

On the other hand, the majority of the assays mirrored from ChEMBL do not contain such explicit bioactivity annotation, but many contain potency specificity. Therefore, users are highly recommended to follow up with results linked under both the bioactivity outcomes and potency (e.g. Active Concentrations) columns. The target-centric page (Figure 3C) provides a summary for the assay experiments associated with a protein target. More importantly, it highlights the compounds exhibiting desired bioactivity for the given target. This service provides users with a path to narrow down chemical modulators with certain potency and follow up with the assay experiments; thus, it can turn into an annotation service for protein and genes. These tools may prove useful for users to explore the biological processes in which a small molecule is involved, or search the chemical modulations of a biological pathway.

Enhanced assay-centric BioActivity Summary tool

As compounds in PubChem are often tested in hundreds or even thousands of bioassays, retrieving data from database and generating a summary view can be time consuming, thus requests to the BioActivity Summary service are put into a queue system. On the other hand, the assay-centric service has recently been improved. It now promptly returns a summary of bioactivity outcome, potency, assay and target information for a single SID or CID input. With this enhancement on data retrieval and presentation, this new tool allows one to quickly gather and export first-hand bioactivity information for a particular compound. One can also bookmark the URL to monitor new discoveries on a known drug or a small molecule of interest.

Dose–response curve drawing tool

Confirmatory assays in PubChem often contain one or multiple series of dose-response data. These assays can be searched using the interface provided on the Entrez PubChem BioAssay 'Limits' page or directly following the URL at http://www.ncbi.nlm.nih.gov/pcassay?term=doseresponse[filt]. To facilitate the assessment of dose–response relationship and better evaluate the potency and efficacy of a tested compound, PubChem now provides a new service to calculate and draw a fitted dose–response curve. In this service, the dose–response data are fitted with the Hill equation based on a nonlinear regression algorithm developed by Pinto *et al.* (13). One can visualize the fitted dose–response curve by clicking on the dose–response icon contained in the assay data table view of a confirmatory assay, or pick an assay accession (AID) and a substance accession (SID) using the web interface at http://pubchem.ncbi.nlm.nih.gov/assay/plot.cgi?plottype=1 (Figure 4).

Table 1. A list of web-based PubChem services for the BioAssay resource

Service	Description	URL example
BioActivity Analysis Services	Home page for bioactivity data analysis services	http://pubchem.ncbi.nlm.nih.gov/assay/
BioAssay Summary	BioAssay summary page for a given AID	http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=myAID
BioAssay Data Table (concise view)	Concise data table for a given AID. The table includes SID,CID, structure, bioactivity outcome, score and active concentration value if available	http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=datatable http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?q=r&aid=myAID
Bioassay Data Table (complete view)	Complete data table for given AID, including all deposited test results	http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=datatable http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?q=r&resultsummary=detail&aid=myAID
BioAssay Test Results Selection	Select/search bioassay test results	http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?q=t&aid=myAID
BioAssay Search	Search BioAssay Database with Entrez	http://www.ncbi.nlm.nih.gov/pcassay/
BioAssay Search, Limits page	An interface for constructing an Entrez query	http://www.ncbi.nlm.nih.gov/pcassay/limits
BioAssay Search, Advanced Page	An interface for reviewing search history and combining search results	http://www.ncbi.nlm.nih.gov/pcassay/advanced
PubChem Deposition Gateway	Chemical structure and bioassay submission tool	http://pubchem.ncbi.nlm.nih.gov/deposit
BioActivity Summary—Assay-centric	BioActivity Summary presented from the assay point of view	http://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.cgi
BioActivity Summary—Compound-centric	BioActivity Summary presented from the compound point of view	http://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.cgi
BioActivity Summary—Target-centric	BioActivity Summary presented from the target point of view	http://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.cgi
BioActivity Summary	BioActivity information for a single SID or CID	http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?sid=mySID http://pubchem.ncbi.nlm.nih.gov/assay/assay/cgi?cid=myCID
Structure–Activity Analysis (SAR)	Structure–Activity relationship analysis and visualization in a heatmap-style display.	http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=heat
Related BioAssays	Related BioAssays by: Activity Overlap, Target Similarity, Deposited Annotation, Same Publication, or Common BioSystems.	http://pubchem.ncbi.nlm.nih.gov/assay/assayHeatmap.cgi?service=assayneighbor&aid=myAID
Scatter Plot/Histogram	BioAssay test results plotting functions	http://pubchem.ncbi.nlm.nih.gov/assay/plot.cgi?plottype=2
Dose–response curve	Draw dose–response curves for confirmatory assays containing dose–response data points	http://pubchem.ncbi.nlm.nih.gov/assay/plot.cgi?plottype=1
Bioassay download	Assay data download service	http://pubchem.ncbi.nlm.nih.gov/assay/assaydownload.cgi

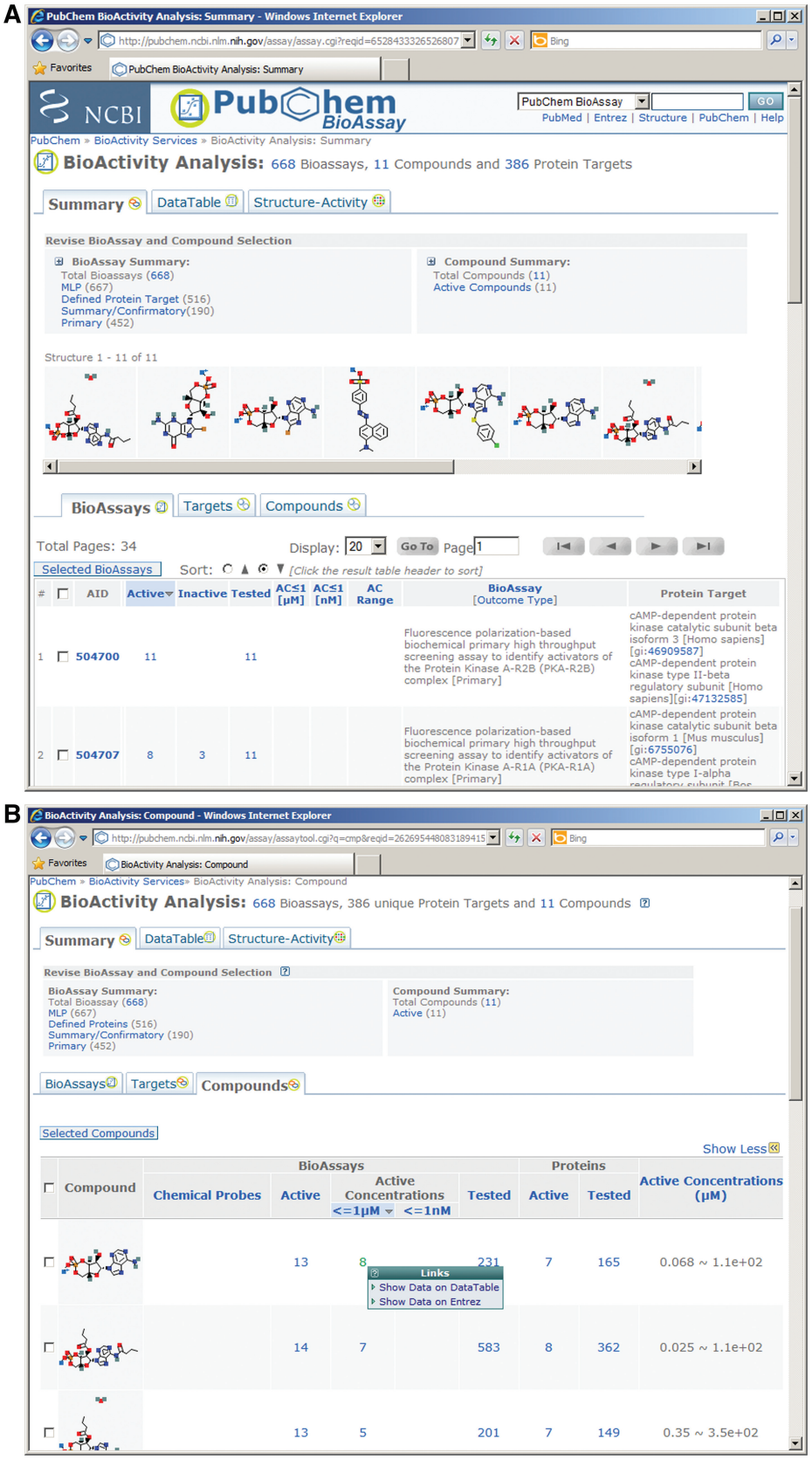

Figure 3. PubChem BioActivity Summary service. (**A**) assay-centric view for multiple compounds; (**B**) compound-centric view; (**C**) target-centric view; (**D**) assay centric view for a single compound.

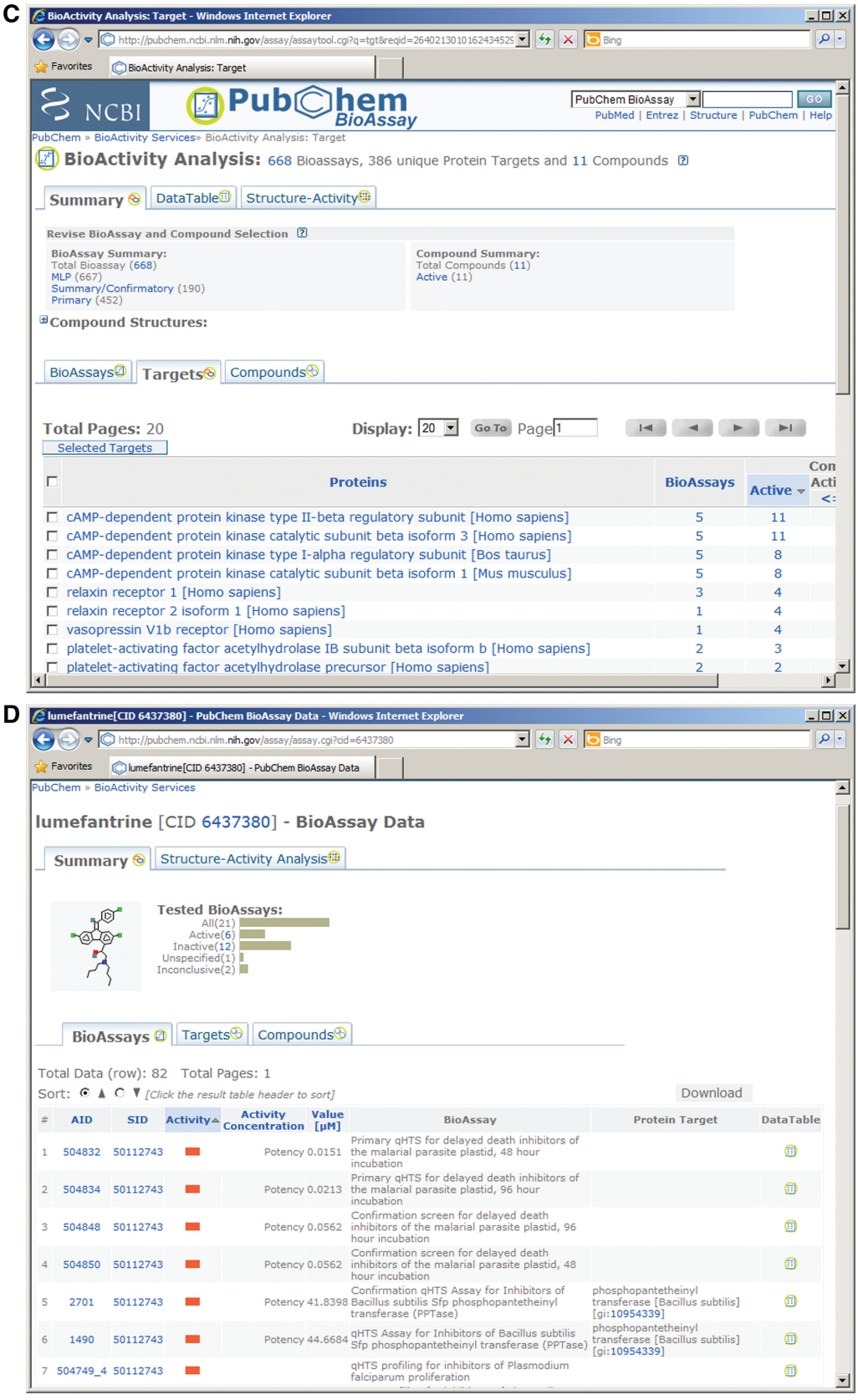

Figure 3. Continued.

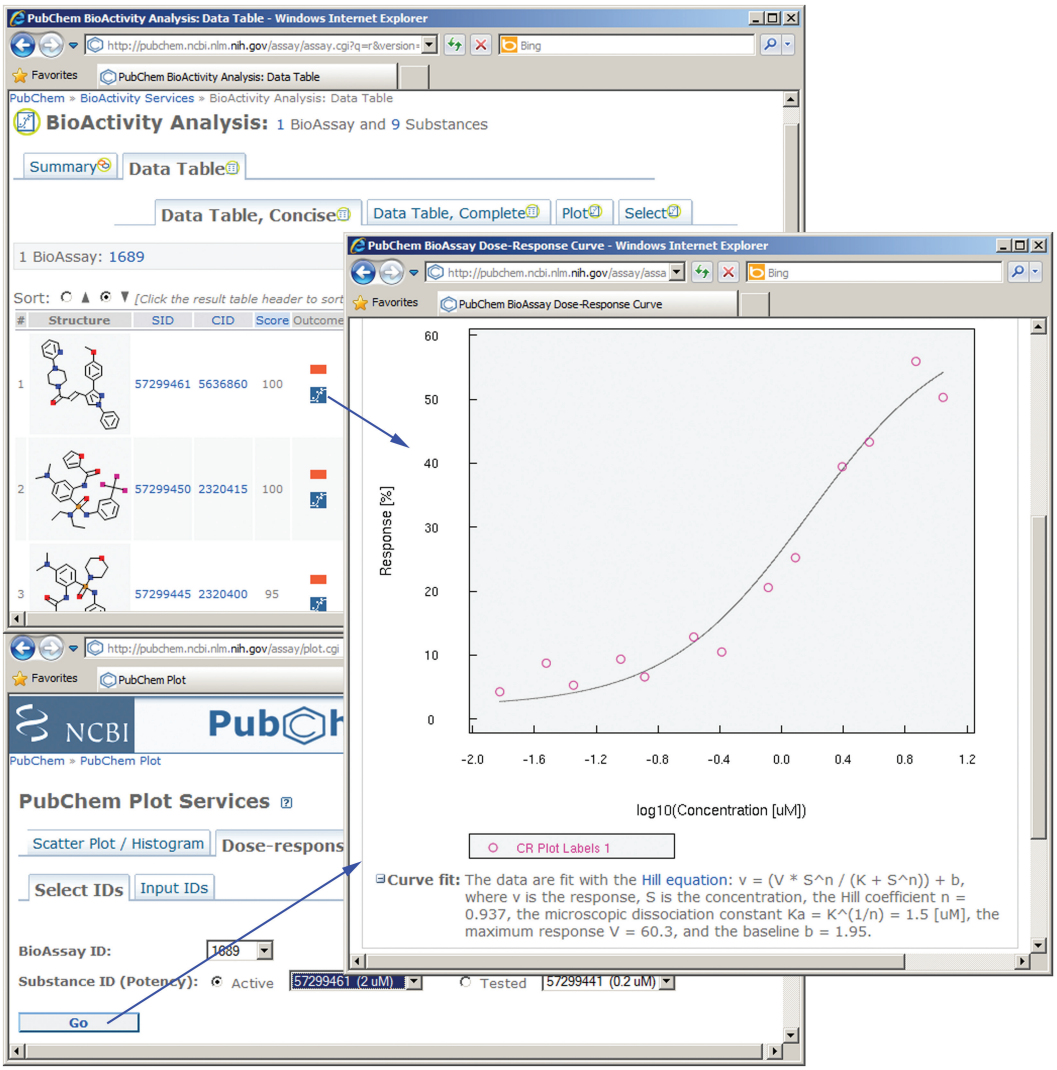

Figure 4. A new service for drawing dose–response curves.

PubChem BioAssay FTP AND DOWNLOAD

PubChem provides multiple services for users to download bioassay records or summarized bioactivity information. Most of the bioactivity analysis services described in this work or previously (2) offer download functionality. Assay descriptions and data table can also be retrieved and downloaded through a programmatic interface using the PubChem PUG/SOAP facilities (http://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html). In this work, we focus on describing two major BioAssay download services: enhanced FTP and a new web service.

BioAssay FTP

PubChem BioAssay FTP (ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay) provides open access to deposited bioassay records. PubChem updates the BioAssay FTP site daily in incremental mode with new and modified bioassay records. One can check the time stamp for the new deposition or update, and check the nature and history of an update by referring to the 'assay.ftpdump.history' file at the FTP site. In addition to depositor-provided bioassay records, annotations derived by PubChem on bioassay relationships can also be downloaded at ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/AssayNeighbors/.

PubChem allows one to download bioassay records in ASN, XML and 'comma-separated values' (**CSV**) formats. The structure of the FTP site is organized according to the respective data formats as shown in Figure 5, e.g. ASN and XML sub-directories provide bioassay records containing both assay description and data in ASN.1 and XML format, respectively. The CSV sub-directory provides CSV-formatted assay data and XML-formatted assay description. The Concise directory contains the XML/ASN/CSV sub-directories with the same structure, but provides only summary assay results including bioactivity outcome, score and active concentration. Due to the large number of bioassay records, bulk downloads from the FTP site are now assisted by the 'zip' compression of multiple records per file with BioAssay AID ranges in the filenames, such as '0000001_0001000.zip'.

BioAssay DOWNLOAD SERVICE

PubChem now provides a web-based interactive Bioassay Download service (Figure 6) to support on-demand bulk download of selected bioassay records. This new web interface can be accessed by following the download icon on a BioAssay Entrez DocSum page (Figure 2) to export records identifies based on a user's search criteria. It can also be accessed directly at http://pubchem.ncbi.nlm.nih.gov/assay/assaydownload.cgi. It allows one to provide a list of AID and retrieve bioassay data and descriptions for up to 1000 bioassay records at a time. One can compose such an AID list by putting together the accessions of assays from a specific data source, or from related assay targets, for example. Moreover, one can also provide a list of PubChem Substance IDs (SIDs) as additional input and request a subset of assay results associated with the specified substances, thus making it straightforward to carry out applications, such as to collect bioactivity data for a group of related chemical structures.

```
|-- ASN
|   |-- 0000001_0001000.zip ...
|-- CSV
|   |-- Data
|   |   |-- 0000001_0001000.zip ...
|   `-- XML
|       |-- 0000001_0001000.zip ...
|-- XML
|   |-- 0000001_0001000.zip ...
|-- Concise
|    |--XML
|    |--CSV
|    |--ASN
|-- README
```

Figure 5. PubChem BioAssay FTP directory structure.

BioAssay DEPOSITION SYSTEM AND UPDATE

The PubChem Deposition Gateway supports chemical and assay data submission through a web-based system at http://pubchem.ncbi.nlm.nih.gov/deposit/. The deposition system also allows bulk data upload via private FTP accounts. A non-trivial task for the system is to validate the submitted data content and provide flexible interface for editing the submissions. Tracking frequent updates to deposited bioassay records represents another challenge as depositors may add additional test results or provide a complete replacement for the entire assay data set.

There is a great variation in complexity of bioassay depositions ranging from very large primary screens with simple endpoints to assays containing dose–response data points or even multiple bioactivity outcomes. Accordingly, the deposition system has been further developed and allows the submission of such information for all types of screening data. The new functionalities include an interface to support the submission of panel assays and categorized comments for organization-specific information. In addition, the user interface of the deposition system is further tailored to better support the submission and the representation of features unique to RNAi data.

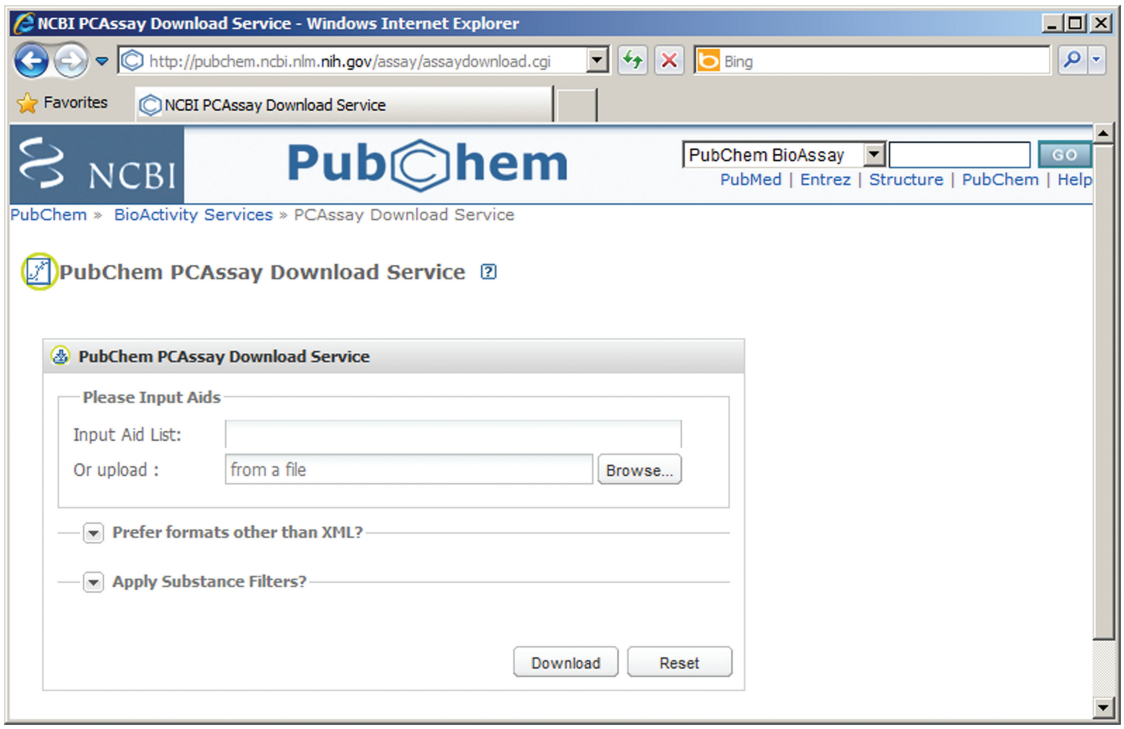

Figure 6. Web-based assay download service.

The depositor is allowed to provide database identifiers in user-defined data columns to expedite the association of gene, nucleotide, literature or other bulk annotation for better reporting and annotating experimental results. Procedures for validating submitted cross-references have also been optimized recently.

Supporting simultaneous submission of such diverse set of data types and sizes from multiple depositors requires interface flexibility and multi-thread processing infrastructure. Many automated (and in some cases manual) checks of incoming data are required to ensure conformity to data specifications and an efficient reporting system is needed for communicating problems within the submitted data to depositors. A number of additional enhancements have been made to the PubChem deposition system to ease and accelerate data submissions. Some of these new additions and improvements to this process are described in this work.

New format for submitting substance records

The PubChem Deposition Gateway now accepts substance submission in CSV format and via web form input. A bioassay test result is always linked to a substance with a unique PubChem substance accession (SID), making it necessary for depositors to submit substance record prior to bioassay data. The majority of substance records in PubChem are small molecules, which have been deposited in the chemistry-standard SDF file format. However, PubChem also accepts non-small molecule substances, in particular, RNAi reagents for submitting RNAi screening results into the BioAssay database.

To ease the submission of substance records for general biologists, substance records can now be uploaded as a standard spreadsheet file, including CSV, or files supported by Excel or OpenOffice. Through the use of PubChem headers in the spreadsheet file, attributes for RNAi reagents can be specified, e.g., cross-references to gene targets, nucleotides and taxonomy records. Small molecule information including structures can also be added to a spreadsheet by using SMILES, synonyms, URLs, external identifiers etc. In addition, a third option for submitting chemical structure is now available via a web form allowing depositors to draw a structure or to generate it from an identifier. Help document is available at http://pubchem.ncbi.nlm.nih.gov/deposit/deposit_help.html#file.

New method for submitting bioassay description

Due to the inherent complexity of bioassay data, support of the deposition system is a demanding task. The deposition system provides a flexible web interface guiding a depositor to enter the assay description information via web forms (or optionally by XML upload) and subsequently upload the assay data table in a CSV format. In addition, one can provide assay description and data in a single XML file via a private FTP account. This method continues to be the most flexible way for an institution to automate the upload of large amounts of data. However, for more occasional users or for those looking for better integration with experimental data files, the simpler spreadsheet format well used by biologists needs to be supported by the deposition system.

A spreadsheet file (CSV, or files supported by Excel and OpenOffice programs) can now be used to fully define a bioassay description. The hierarchical structure of assay description supported by the PubChem BioAssay data model is organized into six groups of information, including general assay descriptions, result definitions, target definitions, external references, categorized comments and panel assay information. The information in each group can be represented with one spreadsheet. In the panel spreadsheet file, each row represents a panel member, and data on the columns provide detailed information for the respective panel member. Details about the format, header standards and accepted data types required for each description group are provided at http://pubchem.ncbi.nlm.nih.gov/deposit/docs/assay_description_csv_tags.html. This new method allows depositors to package the entire definition of the assay into a single Excel or OpenOffice file containing several spreadsheets for each logical section of the assay description. Help documentation for this new assay submission method is available at http://pubchem.ncbi.nlm.nih.gov/deposit/deposit_help.html#assay_descr.ssheetload.

Fast Preview System for quick feedback

An important step for depositing data into PubChem is the ability for depositors to review the content and the presentation of their submission before releasing the data. A 'Preview' facility is provided for both substance and assay depositions. It closely imitates the presentation of the data content in the publication format and can be accessed by depositor for verifying their data integrity and that everything submitted appears as expected. Due to the complexity of bioassay data fields and the corresponding data storage infrastructure, display assay 'preview' was only available for depositors after a few hours delay. A new dataflow and storage scheme has been developed to eliminate the turn-around time and bioassay 'preview' interface is now generated instantly. Multiple preview/update cycles are often required to identify and fix problems before committing the data for publication in PubChem. With this enhancement for the preview facility, depositor can now get instant feedback about their submission through the fast preview page, hence, to address problems without delay. This new feature greatly improves the deposition workflow and allows depositors to complete the bioassay submission and review the presentation of the data set in PubChem within a few minutes.

Stable account ID and ability to change display Data Source

A technical improvement has been made to deposition account management: establishment of stable Data Source Identifier and modifiable Data Source Name (DSN). The former can be thought of as an 'account ID' for each data source which remains unchanged for the life of the account. This stability persists across

personnel changes and is essential for an archiving database like PubChem. The DSN, on the other hand, may be changed by depositors to allow them to adjust their source name attribute visible to the outside world. This change may become necessary when, for example, depositing institution's name changes or the main deposition account holder moves to a different university, etc. As a result, a deposition account ID may be associated with multiple DSN. Tracking source names and source identifiers is very important for PubChem as they can be used as terms in generating Entrez queries.

SUMMARY

PubChem is set up to serve as a public repository for bioactivity data of small molecules and RNAi. An integrated information platform is provided at PubChem with a suite of tools allowing users to query PubChem databases and analyze the retrieved substance records and bioactivity data. PubChem will continue to improve the existing tools and develop new services to optimize the utility of bioactivity data contained in the BioAssay database. Further integration with Entrez system will provide annotation services for genomic resources by linking to small molecule modulators or effective RNAi reagents as identified by screening experiments. A number of new features and data exchange methods have been added to the PubChem Deposition Gateway in the past 2 years to ease and simplify data submission process. An actively undergoing project is to provide a 'lite' version of the PubChem Deposition Gateway with the goal to provide one-time depositors a quick one-step service. PubChem welcomes the community to utilize the resource, provide feedback, and to contribute data content to the repository.

ACKNOWLEDGEMENTS

We thank the ChEMBL team at the European Bioinformatics Institute for their assistance with integrating ChEBML data into PubChem.

FUNDING

The NIH Intramural Research program. Funding for open access charge: US government.

Conflict of interest statement. None declared.

REFERENCES

1. Wang,Y., Bolton,E., Dracheva,S., Karapetyan,K., Shoemaker,B.A., Suzek,T.O., Wang,J., Xiao,J., Zhang,J. and Bryant,S.H. (2010) An overview of the PubChem BioAssay resource. *Nucleic Acids Res.*, **38**, D255–D266.
2. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
3. Bolton,E.E., Wang,Y., Thiessen,P.A. and Bryant,S.H. (2008) PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.*, **4**, 217–241.
4. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Federhen,S. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.
5. Damoiseaux,R., Sherman,S.P., Alva,J.A., Peterson,C. and Pyle,A.D. (2009) Integrated chemical genomics reveals modifiers of survival in human embryonic stem cells. *Stem Cells*, **27**, 533–542.
6. Flockhart,I., Booker,M., Kiger,A., Boutros,M., Armknecht,S., Ramadan,N., Richardson,K., Xu,A., Perrimon,N. and Mathey-Prevot,B. (2006) FlyRNAi: the Drosophila RNAi screening center database. *Nucleic Acids Res.*, **34**, D489–D494.
7. Gamo,F.J., Sanz,L.M., Vidal,J., de Cozar,C., Alvarez,E., Lavandera,J.L., Vanderwall,D.E., Green,D.V., Kumar,V., Hasan,S. *et al.* (2010) Thousands of chemical starting points for antimalarial lead identification. *Nature*, **465**, 305–310.
8. Metz,J.T., Johnson,E.F., Soni,N.B., Merta,P.J., Kifle,L. and Hajduk,P.J. (2011) Navigating the kinome. *Nat. Chem. Biol.*, **7**, 200–202.
9. Lee,B.H., Lee,M.J., Park,S., Oh,D.C., Elsasser,S., Chen,P.C., Gartner,C., Dimova,N., Hanna,J., Gygi,S.P. *et al.* (2010) Enhancement of proteasome activity by a small-molecule inhibitor of USP14. *Nature*, **467**, 179–184.
10. Hirota,T. and Kay,S.A. (2009) High-throughput screening and chemical biology: new approaches for understanding circadian clock mechanisms. *Chem. Biol.*, **16**, 921–927.
11. Hirota,T., Lewis,W.G., Liu,A.C., Lee,J.W., Schultz,P.G. and Kay,S.A. (2008) A chemical biology approach reveals period shortening of the mammalian circadian clock by specific inhibition of GSK-3beta. *Proc. Natl Acad. Sci. USA*, **105**, 20746–20751.
12. Gaulton,A., Bellis,L.J., Bento,A.P., Chambers,J., Davies,M., Hersey,A., Light,Y., McGlinchey,S., Michalovich,D., Al-Lazikani,B. *et al.* (2011) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.
13. Pinto,G.F. and Oestreicher,E.G. (1984) Pocket computer program for fitting the Hill equation. *Comput. Biol. Med.*, **14**, 507–511.